Web Routineness & Limits of Predictability

Investigating Demographic and Behavioral Differences Using Web Tracking Data

Juhi Kulshrestha

Joint work with:

Marcos Oliveira, Orkut Karaçalik, Denis Bonnay, Claudia Wagner

gesis Leibniz-Institut für Sozialwissenschaften

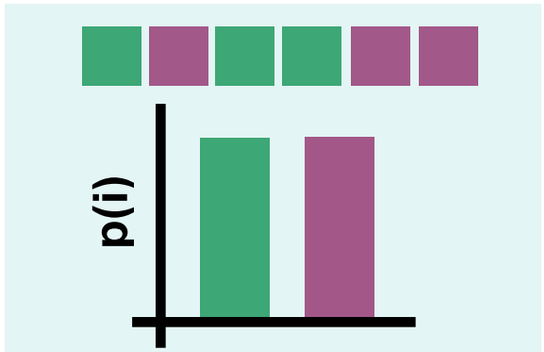
Universität Konstanz

UNIVERSITÄT KOBLENZ · LANDAU

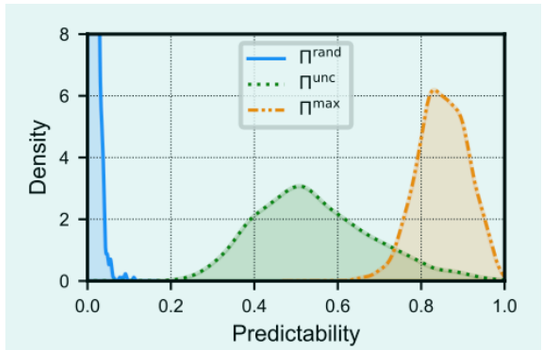
respondi

Université Paris Nanterre

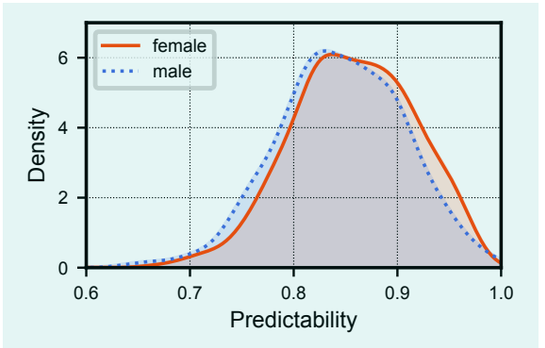
Web routineness & limits of predictability



Predictability measurement framework

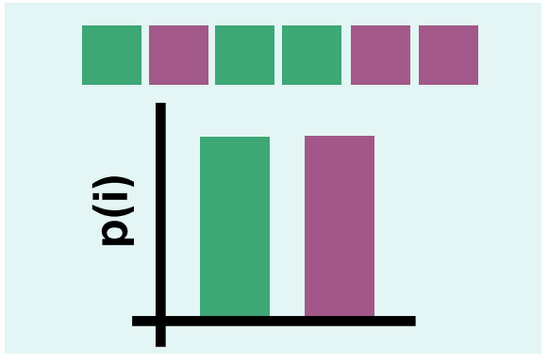


Web routineness and predictability

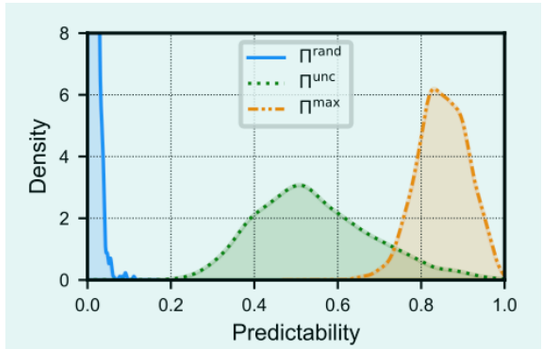


Demographic & behavioral differences

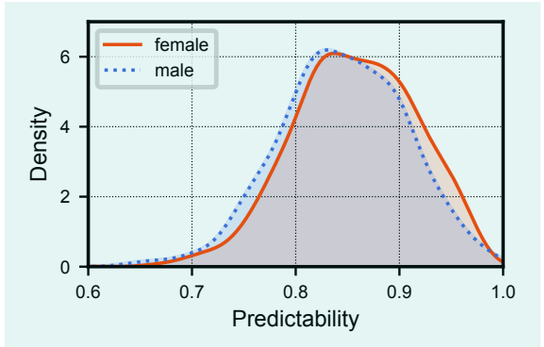
Web routineness & limits of predictability



Predictability measurement framework



Web routineness and predictability



Demographic & behavioral differences

Creating trajectories

- From Web tracking data to trajectories

Time	Duration (s)	URL
2013-05-24 08:19:06	70	https://mail.google.com/mail/u/1/#inbox
2013-05-24 09:15:02	30	https://www.youtube.com/watch?v=I33IGc8Vqy4
2013-05-24 08:35:26	26	https://www.youtube.com/watch?v=EJ6UvCkInRk
2013-05-24 09:15:56	186	https://twitter.com/home

Creating trajectories

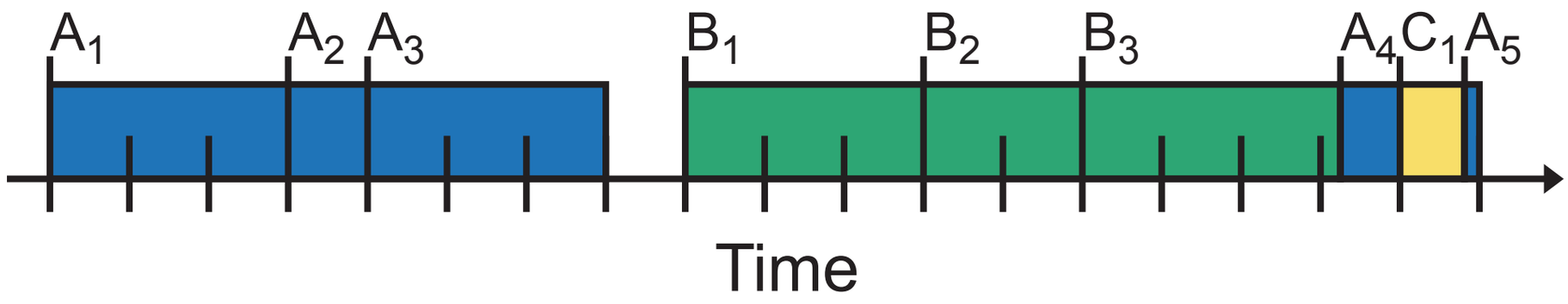
- From Web tracking data to trajectories

Time	Duration (s)	URL
2013-05-24 08:19:06	70	https://mail.google.com/mail/u/1/#inbox
2013-05-24 09:15:02	30	https://www.youtube.com/watch?v=I33IGc8Vqy4
2013-05-24 08:35:26	26	https://www.youtube.com/watch?v=EJ6UvCkInRk
2013-05-24 09:15:56	186	https://twitter.com/home

A trajectory is a discrete sequence of locations visited by a user, where a location can be a website, a domain, or a category.

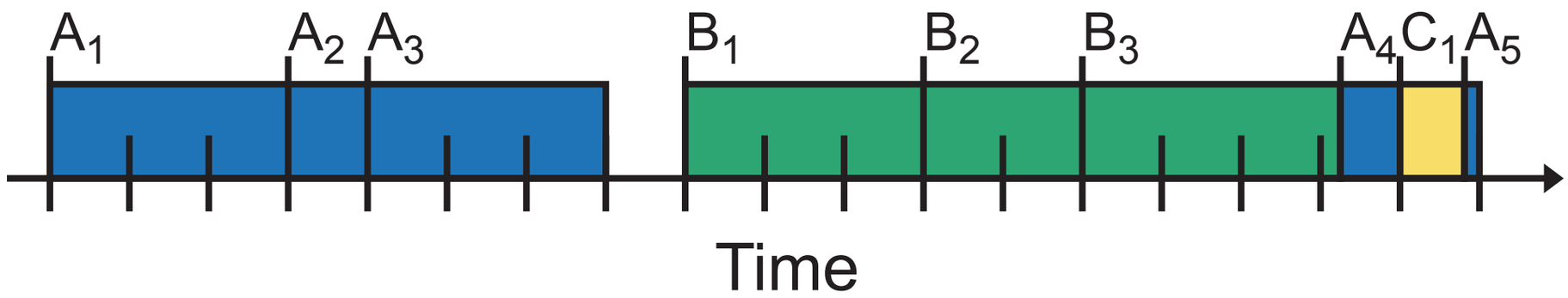
Trajectories

Web tracking data:



Trajectories

Web tracking data:

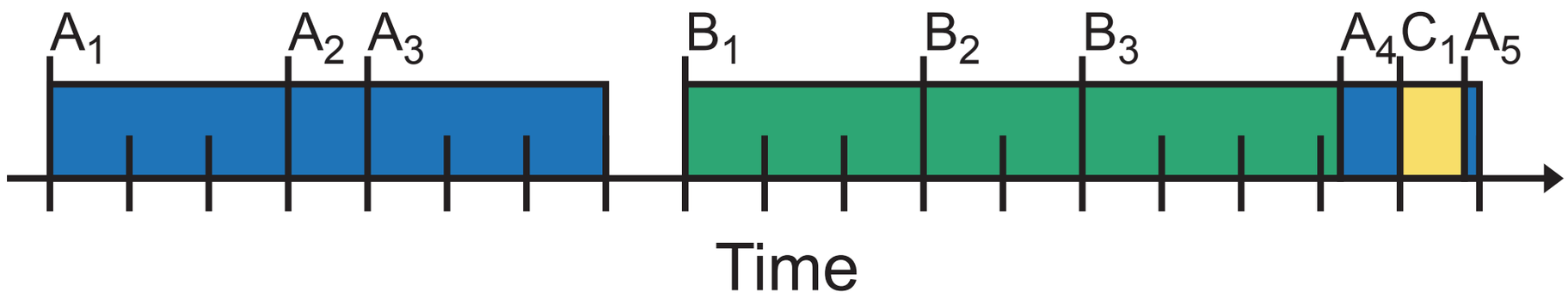


Stationary trajectories *predicting the location visited in the next time step*

$$T^{\text{stat}} = \{A, A, A, A, A, A, A, B, B, B, B, B, B, B, B, A, C\}$$

Trajectories

Web tracking data:



Stationary trajectories *predicting the location visited in the next time step*

$$T^{\text{stat}} = \{A, A, A, A, A, A, A, B, B, B, B, B, B, B, B, A, C\}$$

Non-stationary trajectories *predicting the next visited location*

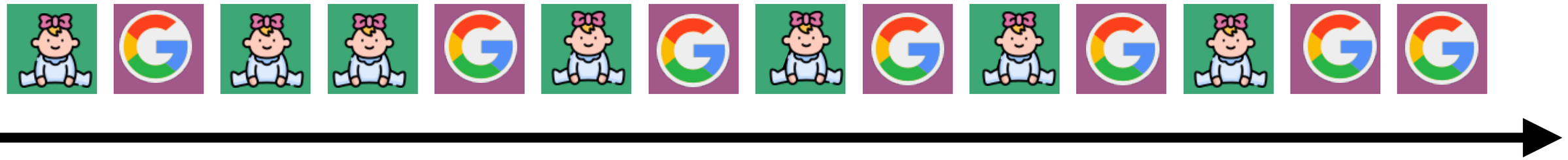
$$T^{\text{binNonStat}} = \{A, B, A, C\} \qquad T^{\text{seqNonStat}} = \{A, B, A, C, A\}$$

Studying trajectories

- We would like to learn about:
  - Location preferences
  - Visitation routines

Studying trajectories

- We would like to learn about:
  - Location preferences
  - Visitation routines



 **= whattoexpect.com/milestones/**

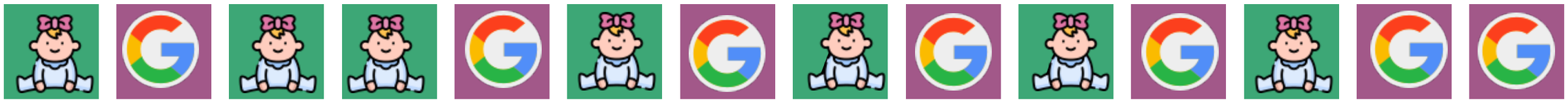
 **= google.com**

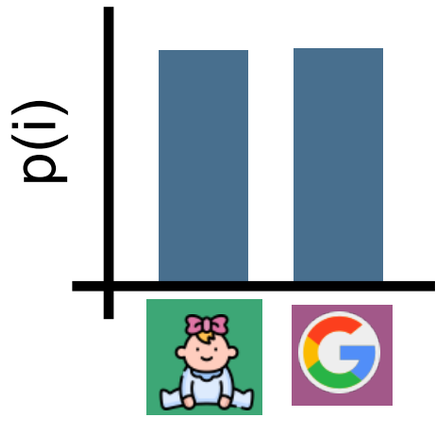
Uncertainty (location preferences)

- Measuring the uncertainty of a random variable.

X

Analyzing the probability distribution of a variable.



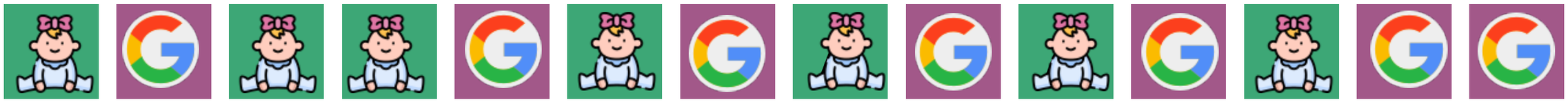


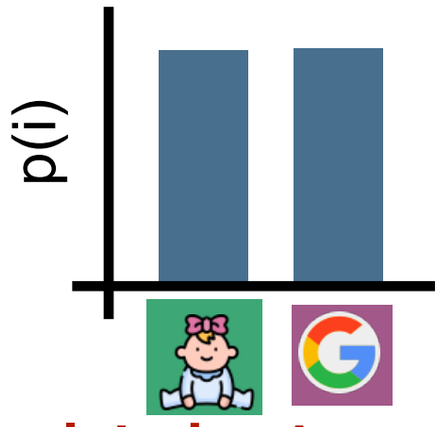
Uncertainty (location preferences)

- Measuring the uncertainty of a random variable.

X

Analyzing the probability distribution of a variable.





Time-uncorrelated entropy (Shannon entropy)

Uncertainty (visitation routines)

- Examining a trajectory as a result of a process

X^L

Analyzing the probability distribution of joint variables.

$$X^2 = X_1 X_2$$

$$X^3 = X_1 X_2 X_3$$

$$X^L = X_1 \dots X_L$$

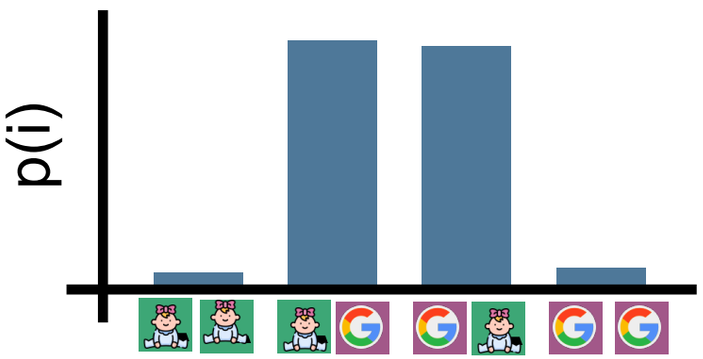
Uncertainty (visitation routines)

- Examining a trajectory as a result of a process

$$X^L$$

Analyzing the probability distribution of joint variables.

For L = 2

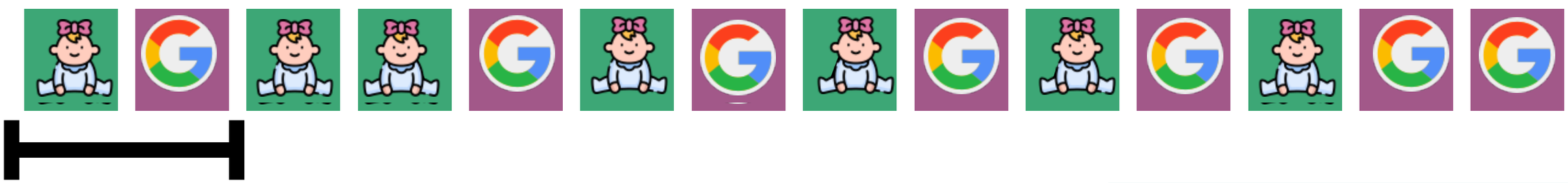


Uncertainty (visitation routines)

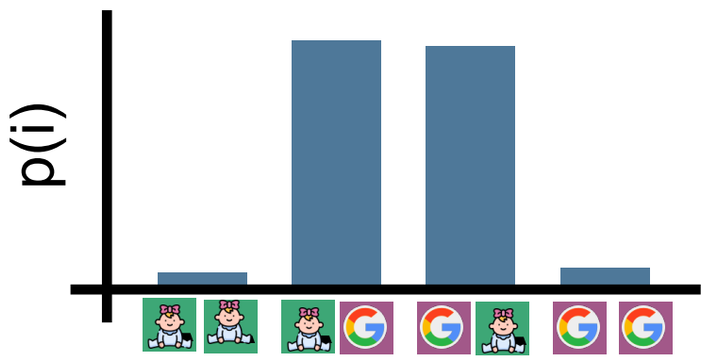
- Examining a trajectory as a result of a process

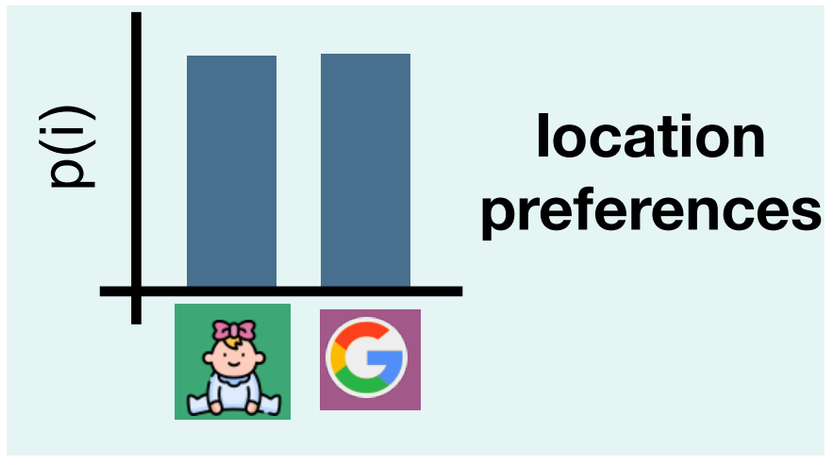
$$X^L$$

Analyzing the probability distribution of joint variables.



For L = 2



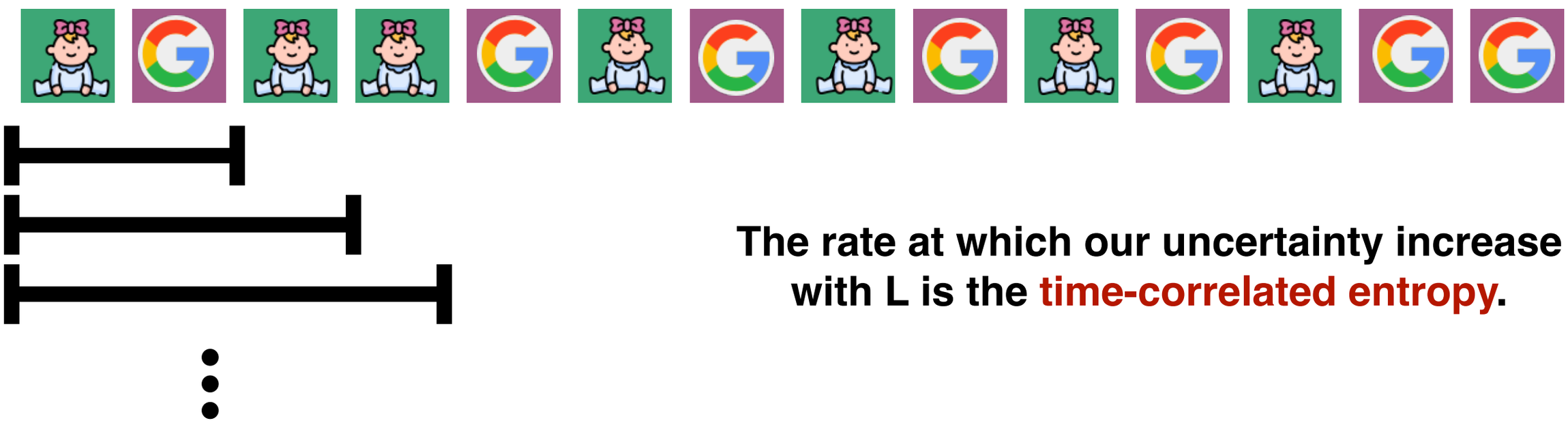


Uncertainty (visitation routines)

- Examining a trajectory as a result of a process

$$X^L$$

Analyzing the probability distribution of joint variables.



The rate at which our uncertainty increase with L is the time-correlated entropy.

From uncertainty to (un)predictability

- What is the probability $\prod$ of correctly predicting future locations given a past series of observation?

From uncertainty to (un)predictability

- What is the probability Π of correctly predicting future locations given a past series of observation?

Limits of Predictability in Human Mobility

Chaoming Song,[1,2] Zehui Qu,[1,2,3] Nicholas Blumm,[1,2] Albert-László Barabási[1,2]*

A range of applications, from predicting the spread of human and electronic viruses to city planning and resource management in mobile communications, depend on our ability to foresee the whereabouts and mobility of individuals, raising a fundamental question: To what degree is human behavior predictable? Here we explore the limits of predictability in human dynamics by studying the mobility patterns of anonymized mobile phone users. By measuring the entropy of each individual's trajectory, we find a 93% potential predictability in user mobility across the whole user base. Despite the significant differences in the travel patterns, we find a remarkable lack of variability in predictability, which is largely independent of the distance users cover on a regular basis.

When it comes to the emerging field of human dynamics, there is a fundamental gap between our intuition and the current modeling paradigms. Indeed, although we rarely perceive any of our actions to be random, from the perspective of an outside observer who is unaware of our motivations and schedule, our activity pattern can easily appear random and unpredictable. Therefore, current models of human activity are fundamentally stochastic (*1*) from Erlang's formula (*2*) used in telephony to Lévy-walk models describing human mobility (*3–7*) and their applications in viral dynamics (*8–10*), queuing models capturing human communication patterns (*11–13*), and models capturing body balancing (*14*) or panic (*15*). Yet the probab
eling framewo
What is the r
havior and to
actions predict

They derived an explicit relationship of the upper-limit with the time-correlated entropy.

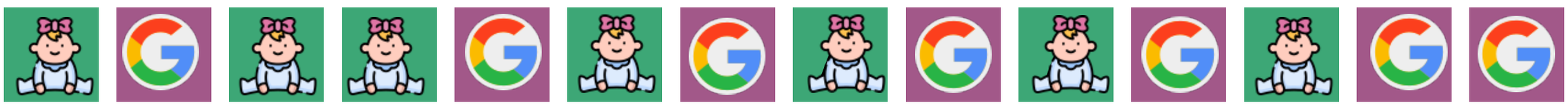
[1]Center for Comple
Biology, and Con
Boston, MA 0211
Medical School, an
Farber Cancer Inst
Computer Science
Science and Technology of China, Chengdu 610054, China.

*To whom correspondence should be addressed. E-mail: alb@neu.edu

19 FEBRUARY 2010 VOL 327 **SCIENCE** www.sciencemag.org

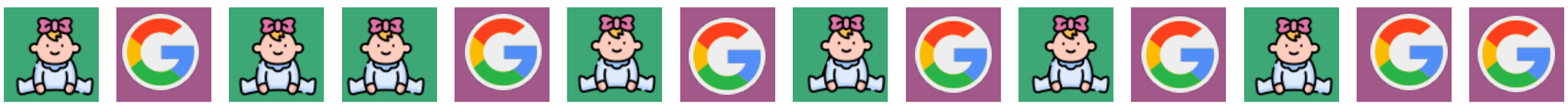
Contextualizing predictability

- Predictability limit which accounts for routines $\Pi^{\max}$



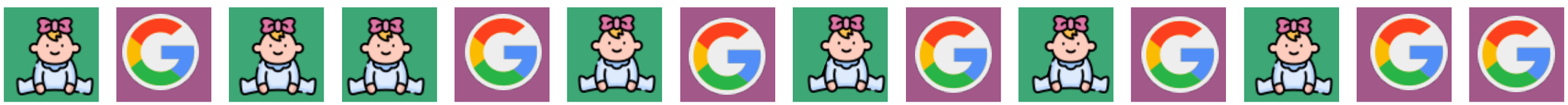
Contextualizing predictability

- Predictability limit which accounts for routines $\Pi^{\max}$
- To understand this value, we create theoretical predictability limits (null models)



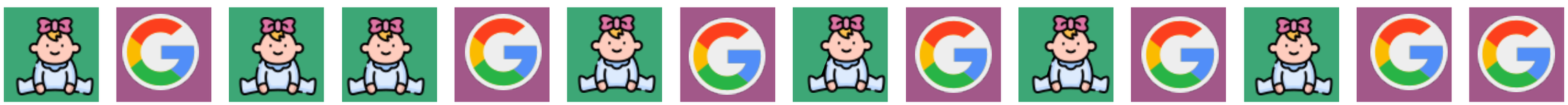
Contextualizing predictability

- Predictability limit which accounts for routines $\Pi^{\max}$
- To understand this value, we create theoretical predictability limits (null models)
  - No preferences: remove all repetition Π^{rand}



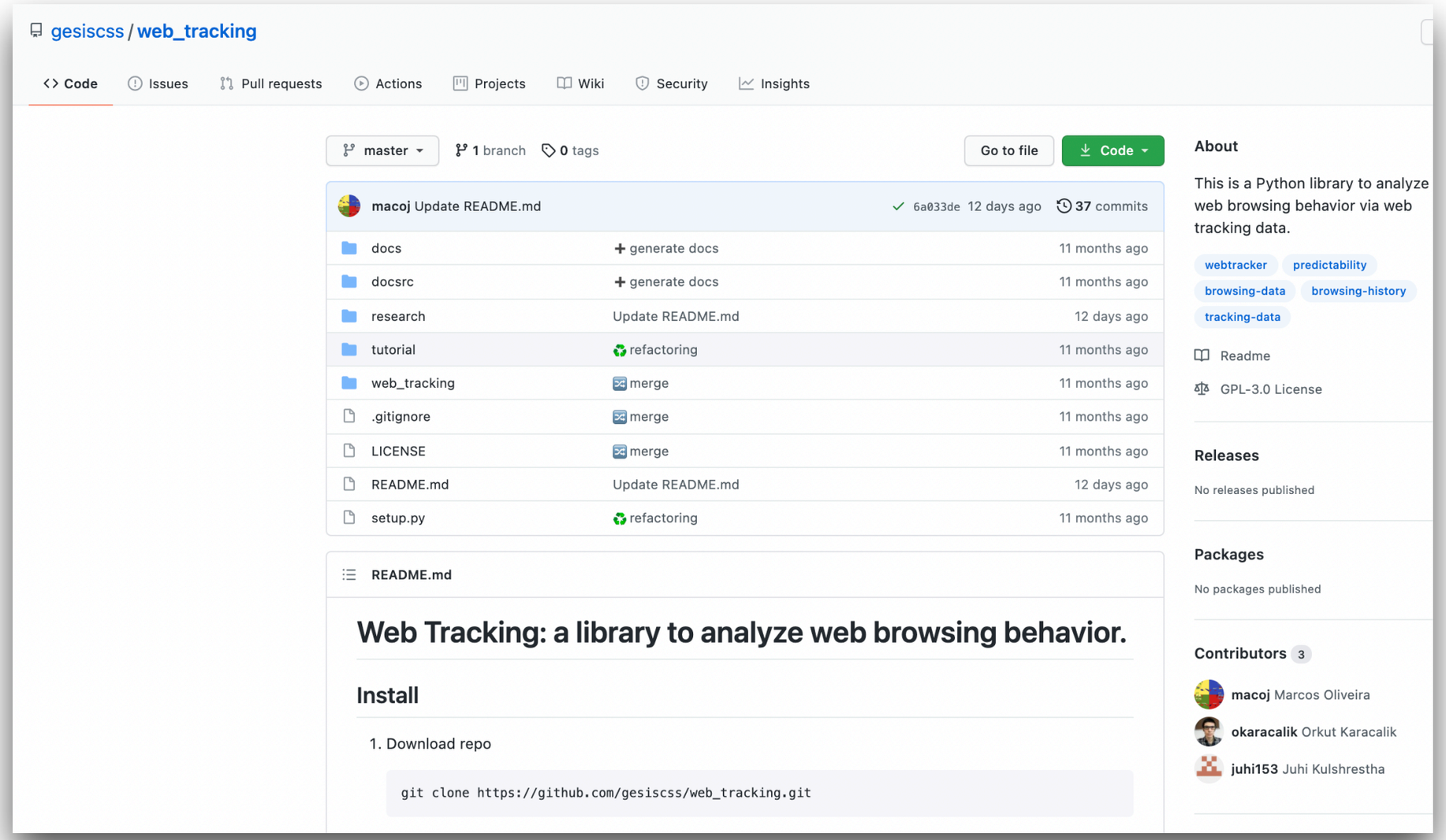
Contextualizing predictability

- Predictability limit which accounts for routines $\Pi^{\max}$
- To understand this value, we create theoretical predictability limits (null models)
  - No preferences: remove all repetition Π^{rand}
  - No routines: shuffle it Π^{unc}

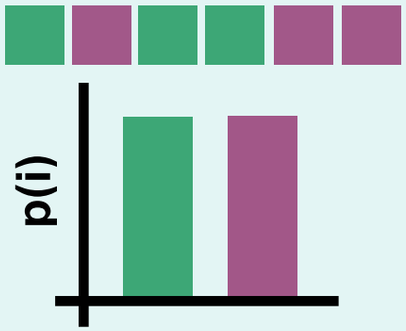


Code available

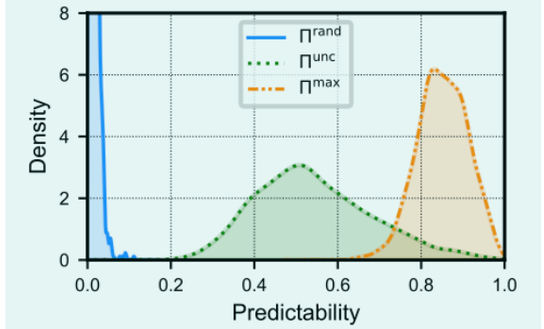
- Python library and notebook tutorials
- https://tinyurl.com/web-tracking-library



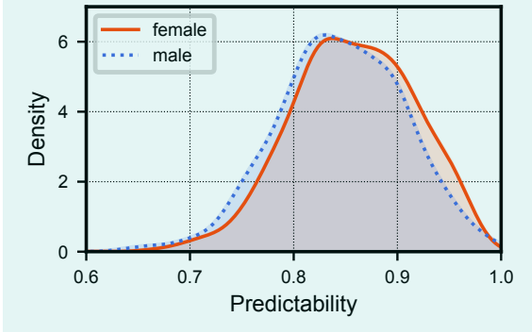
Web routineness & limits of predictability



Predictability measurement framework



Web routineness and predictability



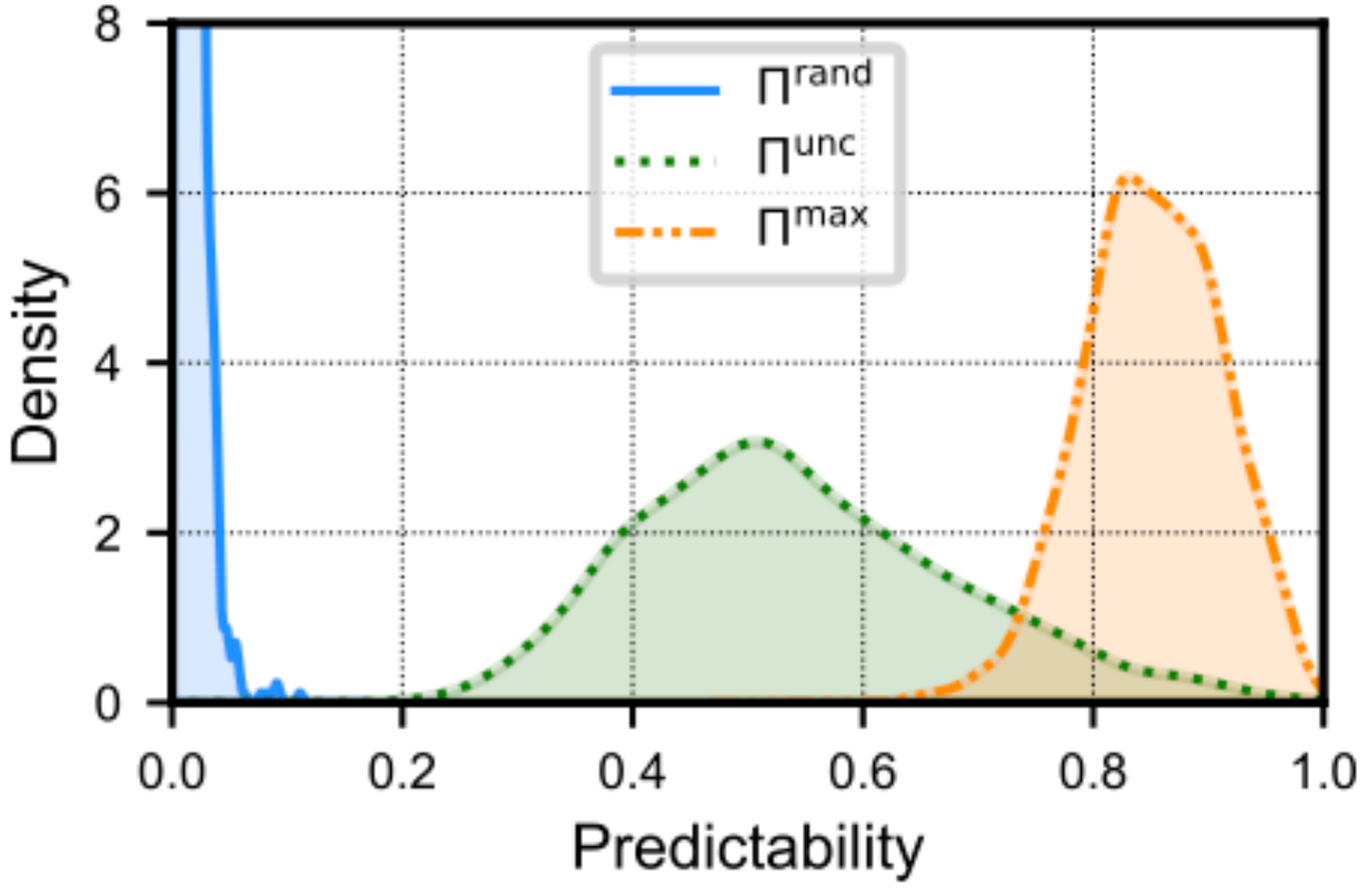
Demographic & behavioral differences

Data

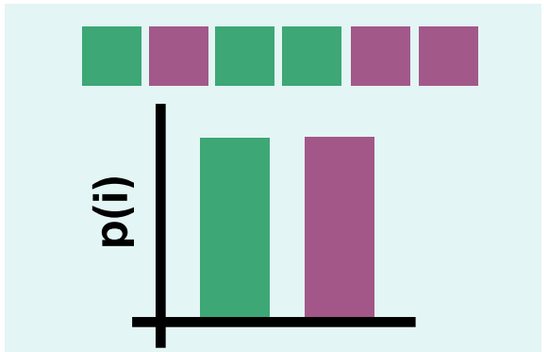
- Sample of German online population:
  - 2,148 individuals.
  - 9,151,243 web visits.
  - 49,918 unique domains.
- Self-reported gender and age.
- Data from a GDPR-compliant digital panel company in Europe.
- https://tinyurl.com/web-tracking-dataset

Predictability of web mobility

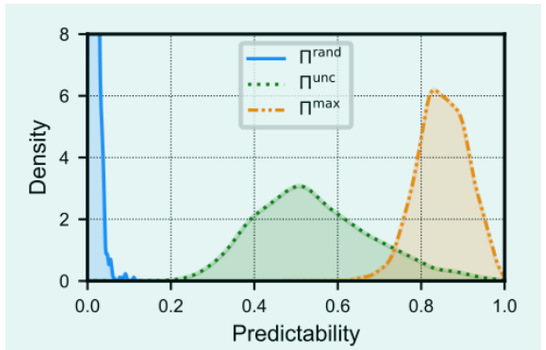
Predict the location visited in the next time step
(stationary trajectory)



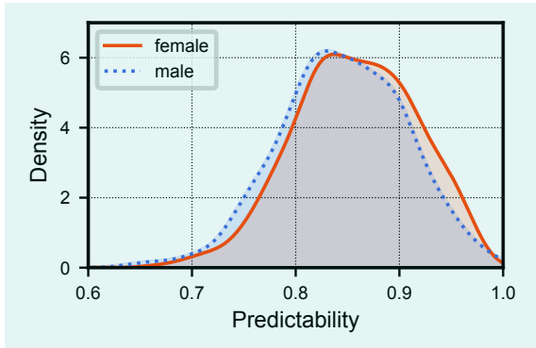
Web routineness & limits of predictability



Predictability measurement framework



Web routineness and predictability



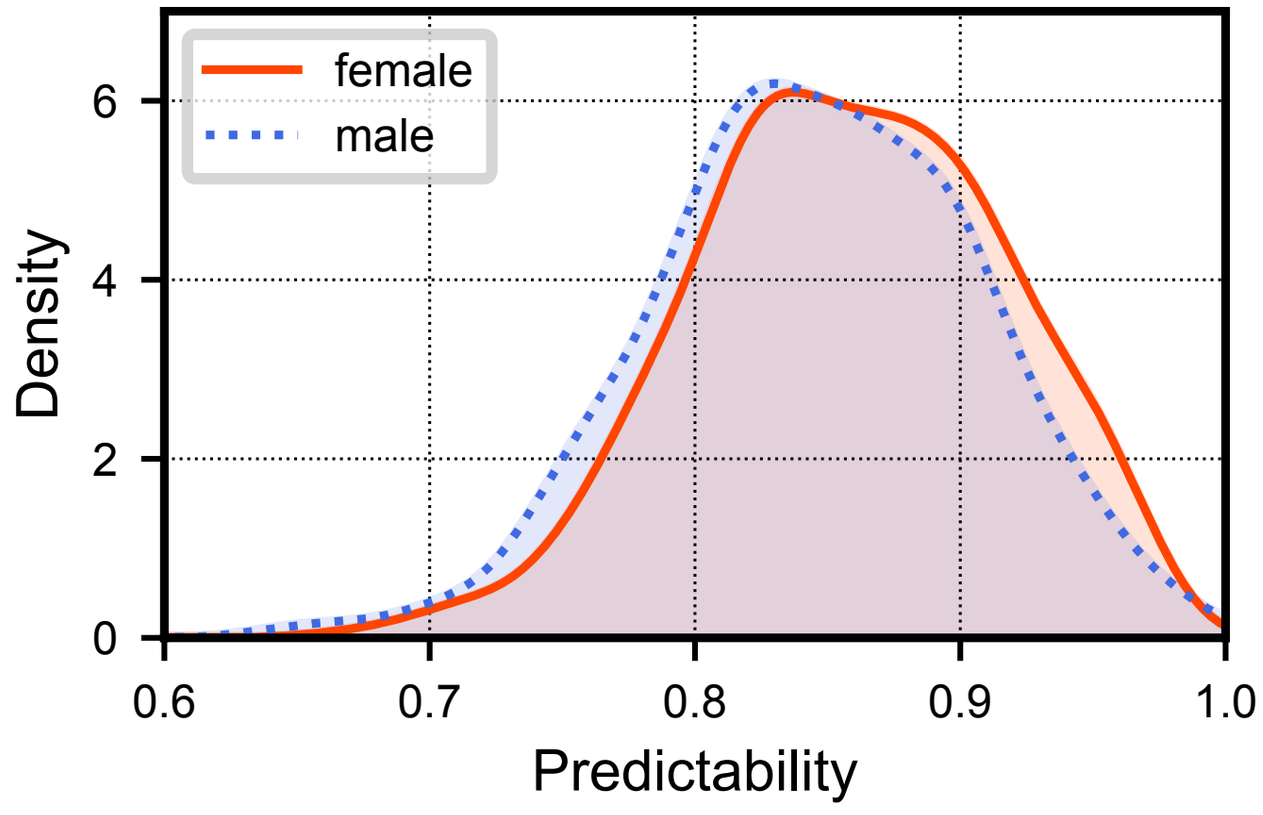
Demographic & behavioral differences

Demographic differences - Gender

- Gender: Male (51.5%) & Female (48.5%)
- Two sample KS test (Hyp: “Two dist are not different”)

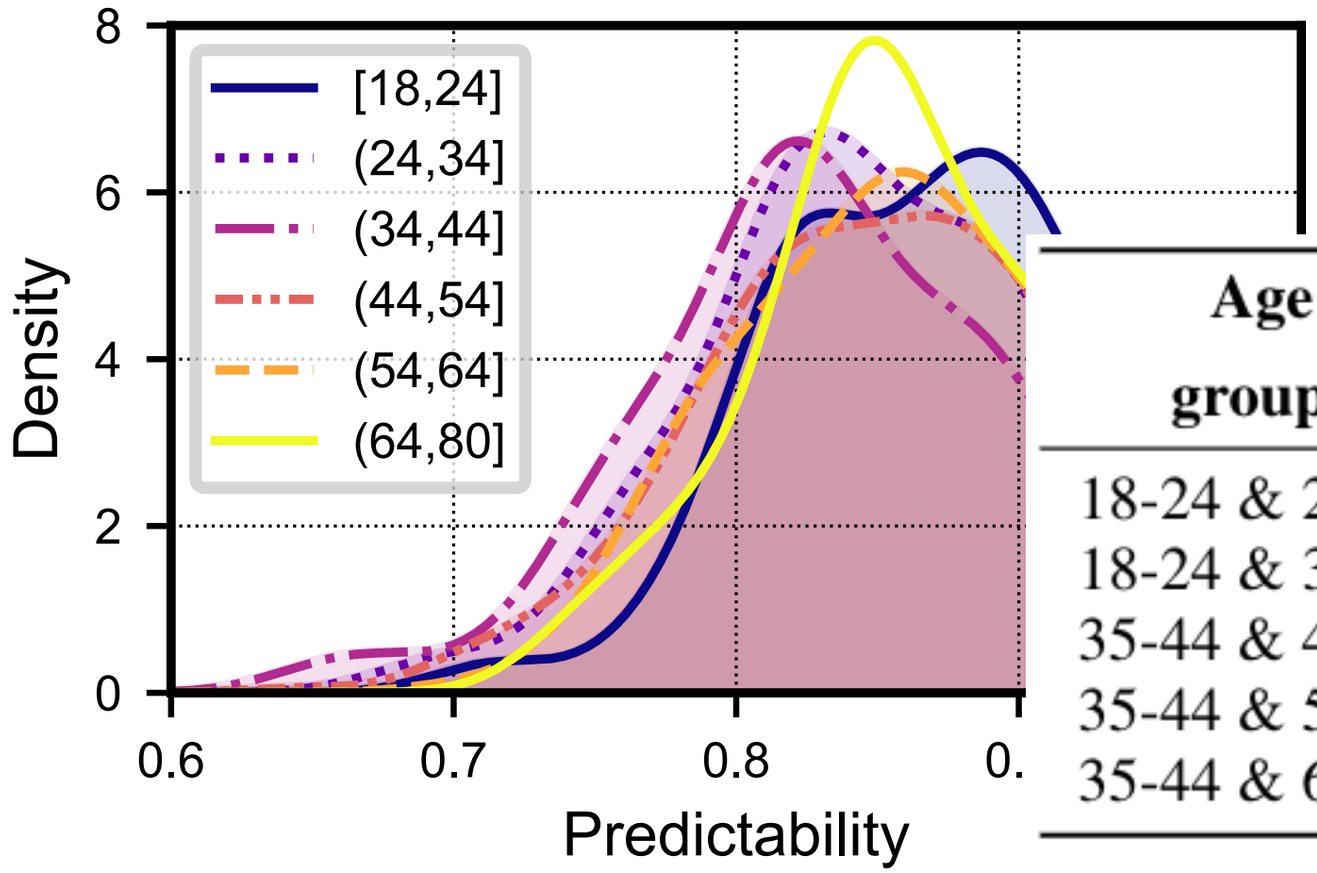
Demographic differences - Gender

- Gender: Male (51.5%) & Female (48.5%)
- Two sample KS test (Hyp: “Two dist are not different”)
  - Rejected hyp. KS Score: 0.086, p-value: 0.008
  - Cliff’s d: 0.11



Demographic differences - Age

- Age (18-24, 25-34, 35-44, 45-54, 55-64, 64-80)
- Two sample KS tests applied pairwise



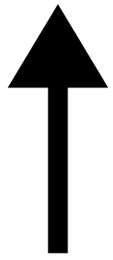
Age groups	KS test		Cliff's delta
	KS score	p-value	
18-24 & 25-34	0.19	8.3×10^{-4}	−0.24
18-24 & 35-44	0.26	1.8×10^{-6}	−0.34
35-44 & 45-54	0.16	6.2×10^{-4}	0.19
35-44 & 55-64	0.21	3.1×10^{-6}	0.25
35-44 & 65-80	0.28	2.1×10^{-4}	0.27

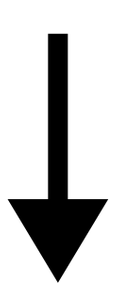
Behavioral differences

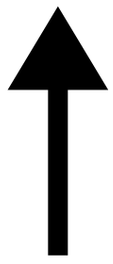
- User activity
  - total time spent browsing
  - total visits
- Diversity of user interests
  - Number of unique domains visited
  - Number of unique categories visited
- User stationarity
  - Mean seconds spent per visit
  - Median seconds spent per visit

Behavioral differences

- User activity
  - total time spent browsing
  - total visits
- Diversity of user interests
  - Number of unique domains visited
  - Number of unique categories visited
- User stationarity
  - Mean seconds spent per visit
  - Median seconds spent per visit



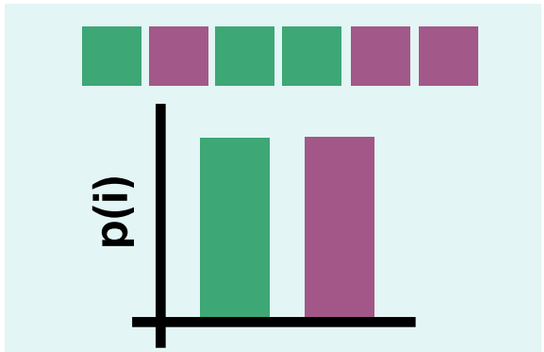




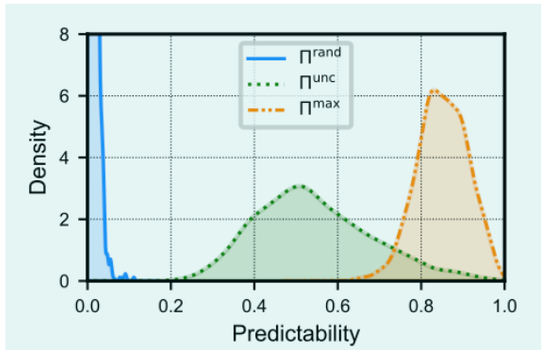
Web Routineness and Limits of Predictability

Investigating Demographic and Behavioral Differences Using Web

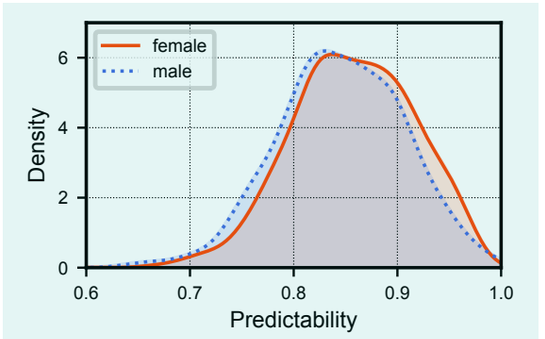
Juhi Kulshrestha, Marcos Oliveira, Orkut Karaçalik, Denis Bonnay, Claudia Wagner



Predictability measurement framework



Web routineness and predictability



Demographic & behavioral differences

Code: https://tinyurl.com/web-tracking-library
Data: https://tinyurl.com/web-tracking-dataset

Thank you!